MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 6: 187–201
DOI 10.3897/mbmg.6.78871

Research Article

Evaluation of non-destructive DNA extraction protocols for insect metabarcoding: gentler and shorter is better

Daniel Marquina[1,2], **Tomas Roslin**[3,4], **Piotr Łukasik**[1,5], **Fredrik Ronquist**[1]

1 *Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden*
2 *Department of Zoology, Stockholm University, Stockholm, Sweden*
3 *Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland*
4 *Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden*
5 *Institute of Environmental Sciences, Jagiellonian University, Krakow, Poland*

Corresponding author: Daniel Marquina (danielmarquinahz@gmail.com)

Academic editor: Florian Leese | **Received** 4 December 2021 | **Accepted** 30 May 2022 | **Published** 16 June 2022

Abstract

DNA metabarcoding can accelerate research on insect diversity, as it is cheap and fast compared to manual sorting and identification. Most metabarcoding protocols require homogenisation of the sample, preventing further work on the specimens. Mild digestion of the tissue by incubation in a lysis buffer has been proposed as an alternative, and, although some mild lysis protocols have already been presented, they have so far not been evaluated against each other. Here, we analyse how two mild lysis buffers (one more aggressive, one gentler in terms of tissue degradation), two different incubation times, and two DNA purification methods (a manual precipitation and an automated protocol) affect the accuracy of retrieving the true composition of mock communities using two mitochondrial markers (COI and 16S). We found that protocol-specific variation in concentration and purity of the DNA extracts produced had little effect on the recovery of species. However, the two lysis treatments differed in quantification of species abundances. Digestion in the gentler buffer and for a shorter time yielded better representation of original sample composition. Digestion in a more aggressive buffer or longer incubation time yielded lower alpha diversity values and increased differences between metabarcoding results and the true species-abundance distribution. We conclude that the details of non-destructive protocols can have a significant effect on metabarcoding performance. A short and mild lysis treatment appears the best choice for recovering the true composition of the sample. This not only improves accuracy, but also comes with a faster processing time than the other treatments.

Key Words

DNA extraction, insects, metabarcoding, non-destructive, taxonomy

Introduction

In the current scenario of global change and dramatic decline in insect biomass and diversity (Hallmann et al. 2017; Van Klink et al. 2020; Outhwaite et al. 2022), we cannot afford to ignore the role of the species that are disappearing and the ecosystem services that they deliver. A first step in addressing this situation is to identify and taxonomically describe as much of the existing diversity as possible. It is well known that insects make up a large proportion of the animal diversity on Earth, but that the essential task of documenting this diversity is far from complete (e.g. Mora et al. 2011). Even in temperate-boreal and well surveyed countries, such as Sweden or Canada, recent large-scale surveys have shown that as much as 20 to 50% of the collected species may be new to science (Hebert et al. 2016; Langor 2019; Ronquist et al. 2019; Karlsson et al. 2020). Nonetheless, when using traditional techniques, even finding the species in mass samples collected by devices such as Malaise or pitfall traps requires an inordinate amount of time and resources. Furthermore, it is also critically dependent on the availability of relevant taxonomic expertise. Therefore, insect diversity researchers are increasingly turning their attention to

Copyright *Daniel Marquina et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

PENSOFT

genetic analysis methods, particularly methods based on high-throughput sequencing (HTS).

Although single-specimen HTS barcoding – the generation of large numbers of DNA barcodes from individual DNA extractions and PCR amplification – is gaining momentum (Srivathsan et al. 2019, 2021), metabarcoding is still the most widely used HTS method for fast assessment of the composition of bulk insect samples that is regarded as fairly reliable. In contrast to single-specimen HTS barcoding, metabarcoding involves the generation of large numbers of DNA barcodes from a sample containing many specimens and a mix of species, without any previous sorting (Taberlet et al. 2012). While popular amongst ecologists (e.g. Buchner et al. 2019; Porter et al. 2019; reviewed in Liu et al. 2019), metabarcoding is less favoured by taxonomists, for the main reason that the vast majority of current metabarcoding protocols involve the homogenisation of the entire bulk sample. This, of course, eliminates any possibility of further investigation of the individuals in the sample, as required for, for example, morphological description of any new species.

One alternative to homogenisation is to use only parts of every specimen in the sample, normally a leg, for DNA extraction (Ji et al. 2013; Beng et al. 2016). However, removing a part of each specimen can take too long to be realistic if the samples are numerous and rich in specimens. Another option is to analyse the DNA released by the specimens into the fixative ethanol of the sample, a technique that has proven successful for bulk samples from freshwater environments (Hajibabaei et al. 2012; Erdozain et al. 2019; Zizka et al. 2019). Unfortunately, for terrestrial insects, which are usually more sclerotised and, thus, leak less DNA passively, metabarcoding of DNA from the preservative ethanol has been found to miss a significant fraction of the insect diversity present in the samples (Linard et al. 2016; Marquina et al. 2019b).

A compromise between tissue homogenisation and analysis of the DNA leaked into preservative ethanol is then to incubate the sample in a digestion buffer that is moderately aggressive, i.e. one with only modest effects on specimen tissues. Such mild lysis methods could potentially retrieve the DNA from the insects efficiently, while preserving the morphological features that are needed to identify or describe the species. Potentially, differences in lysis efficiency could lead to higher representation in the pool of the DNA of small soft-bodied insects relative to homogenisation, facilitating their discovery. They may also allow additional genetic analyses of the specimens at a later stage, if desired. Furthermore, mild lysis protocols may be faster, less labour-intensive, and require fewer steps and less instruments than destructive methods. The composition of such buffers is, with some exceptions, standardised: they consist of a salt (to help precipitate DNA and to separate it from proteins bound to it), a detergent (to break cell membranes and bind to hydrophobic compounds), an inactive pH stabiliser, and a digestion enzyme (proteinase K). Depending on the buffer, it may also contain chelants (compounds that sequester metallic ions that are needed to activate enzymes) or metal salts to activate these enzymes. In addition, the buffers can also contain other compounds that have proteolytic activity without enzymatic intervention, such as dithiothreitol (DTT). It is the presence and concentration of these ingredients that will determine how aggressive the lysis is for the tissue and to what extent it will disrupt the morphological integrity of the specimen.

Mild lysis buffers have often been used in previous metabarcoding studies (e.g. Vesterinen et al. 2016; Ji et al. 2020; Martoni et al. 2021). Yet, to our knowledge, only three studies have investigated whether this is a reliable approach to obtain DNA for metabarcoding. First, Carew et al. (2018) used mock communities made up of macroinvertebrates from freshwater environments, comparing extracts from homogenised samples against extracts obtained with a commercial kit for non-destructive extraction. Nielsen et al. (2019) used mock communities constructed from specimens collected in Malaise traps, comparing the metabarcoding yield from homogenised samples to that from samples extracted in a non-destructive manner. In particular, Nielsen et al. (2019) tested how an increasing complexity of the sample and a larger lysate volume affected the metabarcoding results. More recently, Batovska et al. (2021) tested the use of a commercial non-destructive DNA extraction buffer in recovering low-abundance target species in bulk samples consisting of mixed mock and field-collected communities. All three studies concluded that non-destructive digestion offers a viable solution for metabarcoding and taxonomic work alike: neither Carew et al. (2018) nor Nielsen et al. (2019) found any significant differences in species detection between homogenised and non-homogenised samples, whereas both emphasised the advantage that the non-homogenised samples can be further processed for other purposes. Batovska et al. (2021) did not compare results from non-destructive and destructive DNA extraction methods directly, but noted that they were able to successfully detect their rare target species with non-destructive methods.

Given this consistency in previous outcomes, what is currently missing is a comparison of the performance of different mild lysis protocols. This is the topic we address in this paper. Specifically, we investigate the impact of different parameters of the lysis process (buffer type, digestion time, and purification method) and their effect on the accuracy of metabarcoding in estimating the composition of the original insect sample, as well as on the morphological preservation of the original samples.

Materials and methods

To assess the impact of alternative choices in mild lysis protocols, we tested: 1) two different non-commercial buffers already used in previous studies, one being more chemically aggressive than the other; 2) two different incubation times; and 3) two different ways of purifying

the DNA from the lysate (a manual and a robot-automated process). We then measured the performance of each method by comparing the metabarcoding results to the true composition of mock insect communities. Specifically, we focused on species detection, alpha diversity, and retrieval of the true species-abundance distribution (in terms of individual counts or biomass).

Mock community preparation and DNA barcode reference library

A total of 23 terrestrial arthropod species (including 21 insects, a collembolan and a crustacean) were obtained as live specimens from either standardised cultures of commercial suppliers, donations by other laboratories and the Swedish Museum of Natural History (NRM, Stockholm) vertebrate collection, or collected from the surroundings of the NRM (Suppl. material 7: Table S1). All specimens were killed by submersion in 99% ethanol. Ten types of mock communities (A-J), with four replicates each (a total of 40 tubes), were assembled in 50 mL Falcon tubes (see Suppl. material 1: "Community types" for details). Each community was composed of 22 species, so each of them had a species missing that was present in the rest. In other words, our study explicitly addresses the effect of the extraction protocol on the detectability of species against a community background of standardised complexity.

The number of individuals of each species was always the same regardless of the community type. For example, *Drosophila melanogaster* was represented by six individuals in all ten community types, while *D. yakuba* was represented by three individuals in all community types, except for Community H from which *D. yakuba* was excluded). The total number of insects per community ranged from 70 to 74. Since our specific interest was in the impact of species properties on the detectability of species, rather than of individual variation in body size, the average weight per specimen was computed by recording the dry weight of ten individuals per species. We then selected all individuals of a species from the cultures to be of the same size. As a result of this rationale, the study is well aimed to detect effects of species averages, whereas it provides little information on the effects of added variation in individual size. The proportions of each species in the communities in terms of weight and numbers was registered as abundance in biomass or number, respectively.

Reference DNA barcode sequences were constructed using one additional individual of each species and one individual of the bumblebee *Bombus pascuorum* (used as a control for quantifying index swapping during the sequencing run, see below) as follows. DNA was extracted from all individuals using the KingFisher Cell and Tissue DNA kit on a KingFisher Duo robot (Thermo Fisher Scientific), except for *Encarsia formosa*, *Folsomia candida* and *Tuberculatus annulatus*, which were processed using QIAamp DNA Micro kit (Qiagen) following the manufacturer's protocol. The entire barcoding region of COI (658 bp) and a fragment of 450–490 bp (depending on the species) close to the 5' end of the mitochondrial 16S rRNA gene were amplified and sequenced. COI was amplified using the primers jgLCO1490-jgHCO2198 (Geller et al. 2013), with the exception of *Formica rufa*, which was amplified with the primers LepF1-LepR1 (Hebert et al. 2004) following failed amplification attempts with the previous pair. 16S was amplified with the primers 16Sar-16Sb2 (Simon et al. 1994; Cognato and Vogler 2001) in all species. The PCR mix consisted of one Illustra Hot Start Taq RTG bead (GE Healthcare Life Sciences), 1 µL (10 pmoles) of each primer, 2 µL of DNA template and 21 µL of biology-grade water (final volume: 25 µL); the temperature protocol consisted of an initial phase of denaturation at 95 °C for 5 min followed by 40 cycles of 30 s of denaturation at 95 °C, 45 s of annealing at 50 °C (COI-jg) / 45 °C (COI-Lep) / 48 °C (16S), and 60 s of extension at 68 °C, finishing with a final phase of extension of 10 min at 72 °C. PCR success was checked in an agarose gel and those reactions with positive bands were cleaned of single-strand DNA molecules with ExoSAP-IT (Thermo Fisher Scientific) following the manufacturer's protocol and sent to Macrogen Europe B. V. (Amsterdam, Netherlands) for two-strand Sanger sequencing. The sequences were merged, edited and trimmed of primers using GENEIOUS v8.1.7 (Kearse et al. 2012). The reference sequences of both genes of all 24 species are given in the Suppl. material 2, 3: "COI reference library" and "16S reference library", respectively.

Lysis and DNA extraction

The mock communities were subjected to four different lysis treatments, resulting from the combination of two digestion buffers (referred to as B1 or Gentle, and B2 or Aggressive) and two incubation times (Fig. 1). Buffer B1 was taken from Vesterinen et al. (2016) (which was, in turn, modified from Aljanabi and Martinez (1997)) and consisted of 400 mM NaCl, 10 mM Tris-HCl pH 8.0, 2 mM EDTA pH 8.0, 2% SDS, 0.1% proteinase K 20 mg/mL (which corresponds to 0.069 mM) and molecular biology grade water. Buffer B2 was modified from Gilbert et al. (2007) (see also Ji et al. 2020) and consisted of 115 mM NaCl, 47 mM Tris-HCl pH 8.0, 3 mM $CaCl_2$, 2.4% SDS, 40 mM Dithiothreitol (DTT), 1% proteinase K 20 mg/mL (which corresponds to 0.69 mM) and molecular biology grade water (see Suppl. material 8: "Lysis buffers and purification protocols" for details).

Of the reagents used, sodium dodecyl sulphate (SDS) is a surfactant that breaks cells by disrupting the membrane, and both proteinase K and DTT have proteolytic activity. Thus, having a higher concentration of these compounds, buffer B2 is expected to produce a more aggressive digestion than buffer B1. In addition, buffer B1 contains EDTA, which inactivates proteolytic enzymes, such as nucleases and proteases, while buffer B2 contains Ca^{2+}, which activates these enzymes.

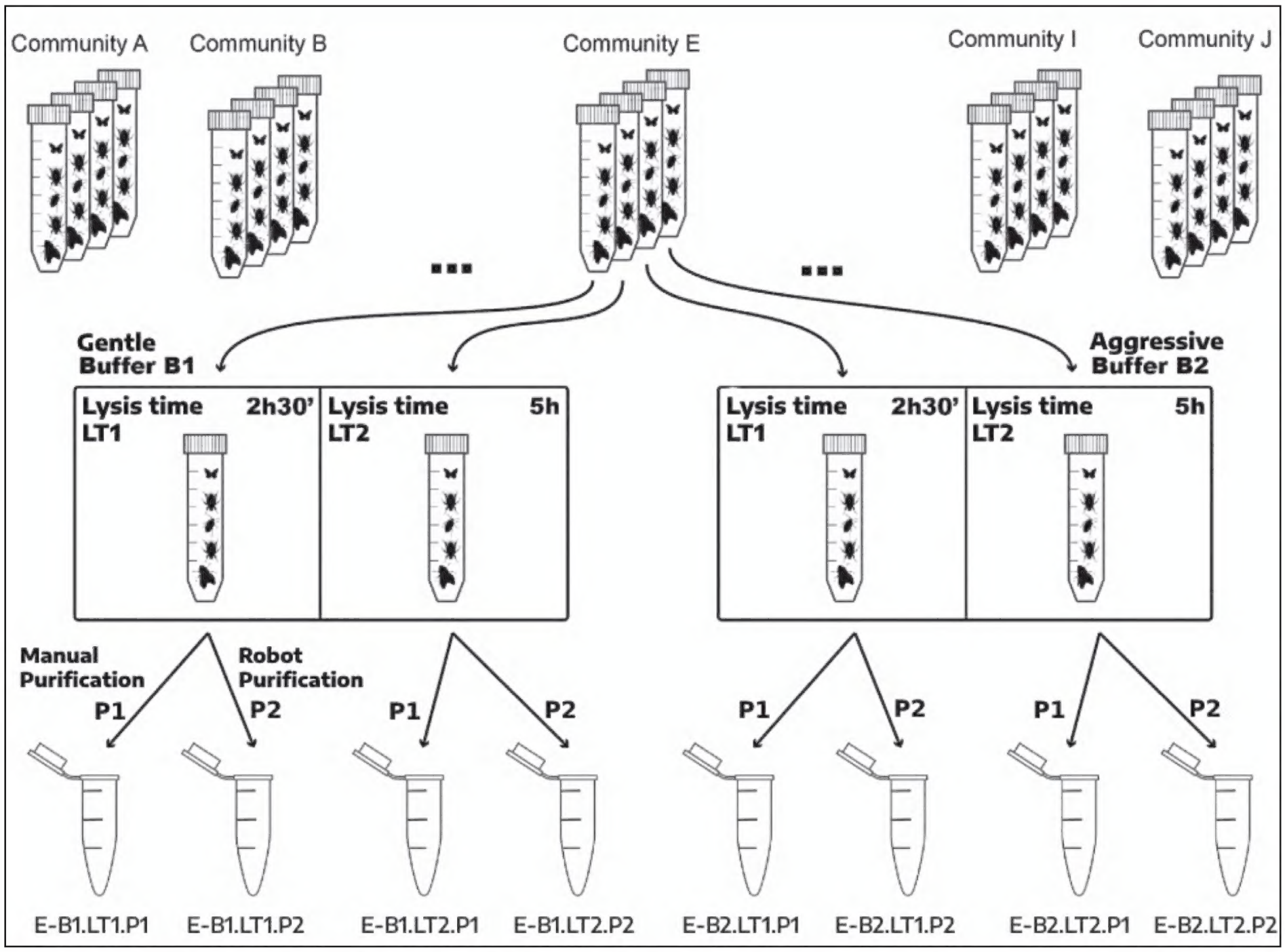

Figure 1. Schematic overview of the experimental design and treatments. Each community (A–J) was represented by four initial replicates (four tubes containing the same mix of species, each of them with the same number of specimens). Each replicate was incubated in one buffer (B1 or B2) for 2 h 30' (LT1) or 5 h (LT2). Then, DNA from each lysate (B1–LT1, B1–LT2, etc.) was extracted by using both a manual salt saturation-salt protocol (P1) or silica-coated magnetic beads in a robot (P2). Thus, for each original community (A–J), eight DNA extracts were obtained, one from each combination of purification method, buffer and lysis time.

In terms of incubation times, samples were split into two different treatments: 2 hours and 30 minutes vs. 5 hours (referred to as LT1 and LT2). When combined with the different lysis buffers, this resulted in four different treatments, each representing a unique combination of buffer and lysis time (B1–LT1, B1–LT2, B2–LT1, B2–LT2). In each case, the samples were incubated in 20 mL of buffer at 56 °C with a slight agitation in an orbital shaker, for each community replicate. Once the incubation time was over, the lysate was decanted out and collected for DNA extraction. The insects were rinsed with molecular biology-grade water, then with clean 70% ethanol, and finally stored in 80% ethanol. The insects remained at all times inside the tubes.

DNA from each lysate was extracted using two purification methods (referred to as P1 and P2). In short, the protocols differed as follows (see Suppl. material 8: "Lysis buffers and purification protocols" for full details). Purification P1 followed a manual salt saturation protocol, modified from Vesterinen et al. (2016) and Aljanabi and Martinez (1997). Taking 7.5 mL of lysate as input material, the proteins and cell membranes were first precipitated using a saturated salt solution, and the DNA was then precipitated using isopropanol, both precipitation steps being aided by centrifugation, with an elution volume of 150 µL. Purification P2 was conducted using silica-coated magnetic beads from the KingFisher Cell and Tissue DNA kit on a KingFisher Duo robot, following the manufacturer's protocol, with an input volume of 225 µL of lysate and an elution volume of 150 µL as well. Summarising, for each community, eight DNA extracts were obtained, making a total of 80 DNA samples (10 mock communities × 2 lysis buffers × 2 incubation times × 2 purification methods). The DNA extracts were measured for concentration and purity (ratio of absorbance at 260 and 280 nm, henceforth referred to as the A260/A280 ratio) using a NanoVue instrument (version 4282 v1.7.3, GE Healthcare Life Sciences). A260/A280 ratios between 1.8 and 2 are considered optimal.

PCR amplification and library preparation

A 321 bp fragment of COI was amplified with a modified version of the primer pair BF2-BR1 (Elbrecht and Leese 2017a, modified in Marquina et al. 2019a) and a 345 bp

fragment of mitochondrial 16S rRNA was amplified with the primer pair Chiar16SF-Chiar16SR (Marquina et al. 2019a). A two-step PCR protocol was followed for library preparation, with a marker-specific first-round PCR with primers with Illumina overhangs attached at the 5' end and an indexing second-round PCR with indexed Illumina adapters. The PCR mix consisted of one Illustra Hot Start Taq RTG bead (GE Healthcare Life Sciences), 10 pmoles of each primer, 2 µL of DNA template and 21 µL of biology-grade water (final volume: 25 µL). The temperature protocol of the two rounds of PCR consisted of an initial phase of denaturation at 95 °C for 5 min followed by 25 cycles of 30 s of denaturation at 95 °C, 45 s of annealing at 48 °C (COI) / 50 °C (16S) and 45 s of extension at 68 °C (first round), or 15 cycles of 30 s of denaturation at 95 °C, 30 s of annealing at 62 °C and 60 s of extension at 72 °C (second round), and a final extension phase of 10 min at 72 °C. Four libraries containing only *Bombus pascuorum* amplicons were added to control for index swapping, and extraction and PCR blanks to control for other sources of error. Without separate index swapping controls, it would be impossible to differentiate between "stray" sequences resulting from tag-switching amongst sequences emanating from the original samples and those resulting from contamination. PCR products were quantified with a Qubit Fluorometer (Thermo Fisher Scientific), pooled in equimolar concentration and purified with QIAquick gel extraction kit (Qiagen) after cutting the bands of the desired length from an agarose gel. They were sequenced on an Illumina MiSeq using v3 chemistry and a 2 × 300 bp paired-end run at SciLifeLab facilities (Stockholm).

Bioinformatic processing

The detailed bioinformatic pipeline with commands and options can be found in Suppl. material 4: "Bioinformatic pipeline". Sequences were processed with the OBITools pipeline (Boyer et al. 2016), complemented by other programmes and scripts (mainly from https://github.com/metabarpark/R_scripts_metabarpark). An attribute containing sample information was added to all sequence headers before pooling together all the libraries in a single file (COI and 16S separately). Reads were quality checked with FastQC (Andrews 2010), the ends trimmed when average quality dropped below a Phred score of 28, and finally paired-end-merged, discarding sequences with an alignment score lower than 40. Primers were trimmed away using CUTADAPT v1.8.0 (Martin 2011), and only the reads with the desired length were kept for downstream analysis (310–330 bp for COI, 290–370 bp for 16S).

Subsequently, the reads were dereplicated and chimeras were filtered out using the *uchime_denovo* function in VSEARCH v2.7.1 (Rognes et al. 2016) and clustered into Molecular Operational Taxonomic Units (MOTUs) using SWARM v2.1.13 (Mahé et al. 2015). The maximum distance *d* allowed during clustering was 13 for COI and 6 for 16S, which corresponds to a sequence divergence of 3–4% and 1–2% for COI and 16S, respectively. MOTU occurrence tables were curated with LULU v0.1.0 (Frøslev et al. 2017) to collapse NUMT-derived erroneously generated MOTUs into their parent MOTUs. The centroid sequences of every MOTU generated by SWARM were compared against the previously constructed reference libraries using the ecotag script from OBITools, and the resulting file was merged with the abundance table. The resulting file was curated with the refine_MOTU_table script (https://github.com/metagusano/metabarcoding_scripts) to remove MOTUs with a relative abundance per sample lower or equal to that generated by index swapping (i.e. abundance of reads assigned to *Bombus pascuorum* in the mock community samples), as well as with less than 10 reads in total, and to collapse MOTUs with coincident species identification.

Statistical analysis

All data analyses were performed in R v.3.3.3 (R Core Team 2017). We first visualised the differences in the estimated community composition produced by different methods, using non-metric multidimensional scaling (NMDS), based on Bray-Curtis dissimilarity (functions *vegdist* and *metaMDS* from package 'vegan' (Oksanen et al. 2013)). To then pinpoint the effects of individual methodological choices, we ran a series of tests. First, we tested whether differences in digestion buffer and incubation time during lysis, and in purification method post-lysis, had effects on the concentration and purity of the DNA extract. Specifically, we applied a split-split-plot analysis of variance (ANOVA) to values of DNA concentration as a function of Buffer (main plot), Incubation time (subplot) and Purification (sub-subplot), with Community type (A-J) as replicate (function *ssp.plot* from package 'agricolae' (de Mendiburu 2019)). For the purity of the DNA extract, another split-split-plot ANOVA was fitted to the A260/A280 values with the same formula as for the concentration. Subsequently, a post-hoc least significant difference test (function *LSD.test* from package 'agricolae') was used to conduct pairwise comparisons between groups. Then, to investigate if the differences in concentration and purity of the DNA extract had any effects on the recovery of species after sequencing, we repeated the split-split-plot analysis with the number of species recovered as the response variable. This analysis was performed for each marker separately.

As the experimental set-up did not allow for an analysis of correlation of real abundance to read abundances, we then investigated how the lysis buffers and the digestion times performed in recovering compositional-related metrics from the samples. We compared the differential in alpha diversity (Shannon index, H') from the original mock communities to estimates obtained through metabarcoding (function *diversity* from package 'vegan'). We also calculated the Kullback-Leibler Divergence between the observed community and the original, known composition of the mock community (function *KLD* from

package 'LaplacesDemon' (Statisticat 2018)). In brief, the Kullback-Leibler Divergence takes a distribution as its reference (in this case, the true proportions of the species in the mock communities) and calculates how much information must be added to a second distribution (the proportion of reads of each species in the metabarcoding sample) to make it equal to the reference. To model the impact of methodological choices on the Shannon Index and the Kullback-Leibler Divergence between the true and observed sample, we applied another split-split-plot analysis with the values as a function of the three factors, as previously done for purity and concentration. The purification term was not significant in any of the split-split-plot analyses of the Shannon Index and the Kullback-Leibler Divergence and was thus dropped. These analyses were done separately for COI and 16S and using both reference distributions based on relative abundance and based on relative biomass, and with both datasets rarefied to the number of reads of the sample with the lowest number of reads.

Results

Regardless of the digestion treatment, insects were recovered in a good state and maintained exoskeletal integrity as well as colour features. We observed no effect of incubation time, but those insects from communities digested with buffer B2 presented a faint red-brownish tone and a slightly higher transparency after the lysis step (Fig. 2).

The MiSeq run produced a total of 5,112,064 sequences of COI and 9,700,315 of 16S, of which 4,467,494 (reads/sample = 16,682 ± 14,797 (mean ± s.d)) and 9,264,057 (reads/sample = 35,710 ± 23,509 (mean ± s.d)), respectively, passed the quality filters. With COI metabarcoding, we recovered all 23 species, but no reads were obtained from sample C2.2.1, so this sample was excluded from all subsequent analyses. With 16S, we did not recover *Porcellionides pruinosus* nor any of the *Formica* species. MOTU tables with species identification, abundance in each sample and representative sequences are provided in the Suppl. material 5, 6: "COI MOTU table" and "16S MOTU table", respectively. All extraction and PCR blanks featured some reads of almost all species (see Suppl. materials for details). The read numbers roughly corresponded to those associated with unintended tag combinations (i.e. combinations likely generated by tag switching) present in the samples named E1–E7 and were thus disregarded as likely products of index jumping (Kircher et al. 2012).

DNA concentration and purity

The concentration of the DNA extracts ranged from 4.6 to 371.5 ng/μL (Suppl. material 7: Table S2). The lowest mean values corresponded to the replicates incubated in buffer B1 (for 2 h 30' or 5 h) and subsequently purified using the extraction robot. By comparison, the highest mean values corresponded to the replicates incubated in buffer B2 for 5 hours, then purified with the manual salt saturation method. The ratio A260/A280 ranged from 1.05 to 2.00. The lowest mean values corresponded to the replicates incubated in buffer B1 for 2 h 30' and purified with the robot and those incubated in buffer B2 also for 2 h 30' and purified manually. The highest mean values

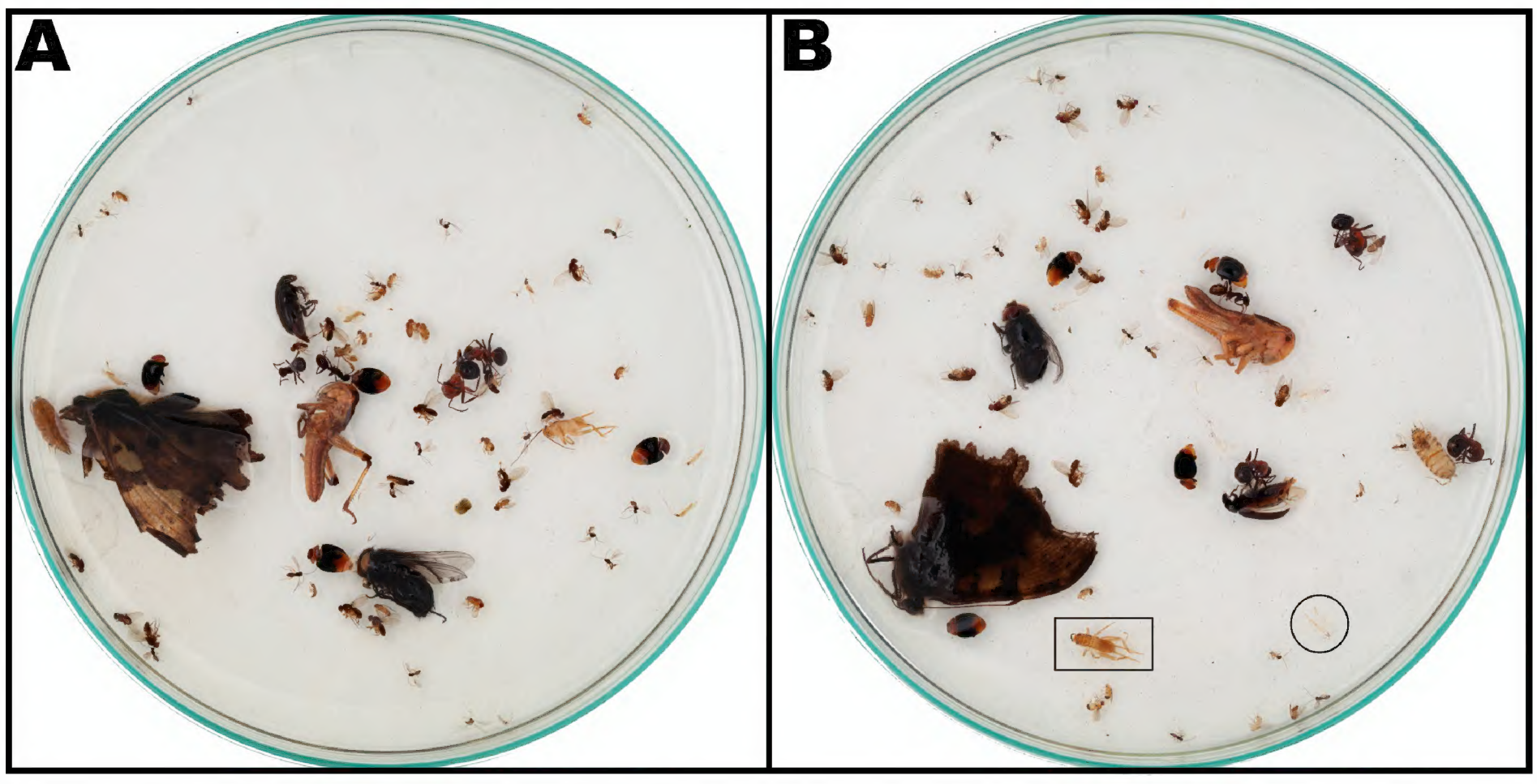

Figure 2. Examples of the mock communities after digestion. Insects in the community type E incubated in lysis buffer B2 for 5 h (**A**) are slightly discoloured (effect of the storage in ethanol), but the morphology and the colouration patterns are well preserved. Insects in the community type E incubated in lysis buffer B2 for 5 h (**B**) have a faint reddish tone (*Acheta domesticus* specimen inside the rectangle) and the colour of some of the small individuals have slightly faded (*Aphidoletes aphidimyza* specimen inside the circle), but the colours and morphology of most specimens is still reasonably well preserved.

corresponded to the replicates incubated in buffer B1 for 5 hours and purified manually.

All three factors and their interaction had significant effects on the DNA concentration of the extracts (Suppl. material 7: Table S3). Lysis with buffer B2 increased the DNA concentration, same as longer (LT2) incubation (Fig. 3, upper panel). In addition, extracting the DNA from the lysate with the salt saturation method generated extracts with a much higher concentration than those extracted with the automated robot extraction using silica beads. Undoubtedly, this was partly influenced by the starting volume being 7.5 mL in the first case and only 0.225 mL in the latter.

Several factors had a significant effect on the purity of the extract (ratio A260/A280; Suppl. material 7: Table S4). However, neither Buffer nor the three-way interaction showed significant effects. When buffer B1 was used for lysis, the purification method based on salt saturation produced extracts of significantly higher purity than did the method based on silica beads, but the incubation time did not have a strong effect. In contrast, when using buffer B2, LT2 improved the purity of the extract (Fig. 3, lower panel).

These differences had no significant effect on the number of species recovered with 16S. However, for COI, the Buffer and Purification effects were both significant, albeit small. Specifically, the number of species recovered was slightly higher for buffer B2 and for the salt saturation protocol. For COI, the buffer affected the mean number of species recovered (Suppl. material 7: Table S5). Against this background, the lysis time did not modify the mean (Suppl. material 7: Table S3), but the purification method did (Suppl. material 7: Table S3). Given significant two-way interactions between Buffer and Purification, we arrive at a scenario where combination B2–P1 provides the highest response in terms of species recovered, with B2–P2 second, then B1–P1 and B1–P2 the lowest (Suppl. material 7: Fig. S1). For 16S, there was no significant effect of any of the factors in the number of species recovered nor of their interactions (Suppl. material 7: Table S6). The average number of species recovered ranged between 20.5–21.5 for COI and 18.2–19 for 16S.

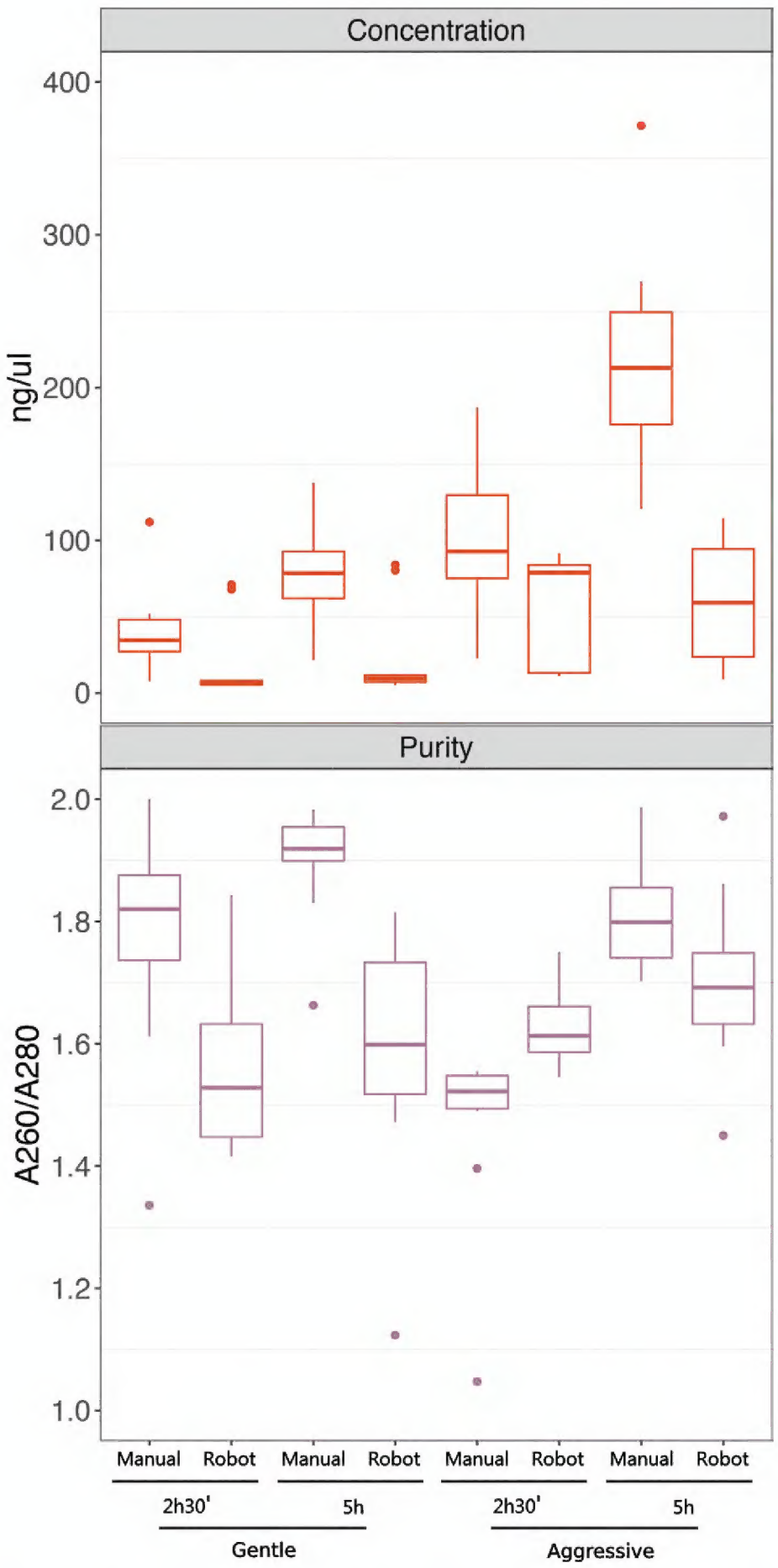

Figure 3. Concentration and purity of the DNA extracts from different extraction methods. DNA concentration (upper panel) clearly increases with buffer aggressiveness and incubation time using the manual salt saturation purification protocol, while the increase due to incubation time is less clear, but the effect of lysis buffer can still be appreciated when using the robot purification protocol. Note that the starting input volume of lysate is 7.5 mL for the manual purification method and 225 µL (30 times smaller approximately) for the robot, while elution volume is 150 µL in both cases. Purity of the DNA extract (lower panel) is higher for the manual purification and the longer incubation times, regardless of the lysis buffer.

Correspondence between mock community and metabarcoding results

In terms of community composition, the metabarcoding results did not resemble the mock communities in neither specimen abundance nor biomass and they all clustered together in the NMDS plot (Suppl. material 7: Figs S2–S4). Although more similar amongst themselves than to the mock communities, replicates treated with different lysis buffers and incubation times showed great variability. Purification replicates, on the other hand, were very similar to each other (see bars distribution in Suppl. material 7: Fig. S3, S4). A few large species dominated most samples, namely *Calliphora vomitoria* (Diptera), *Acheta domesticus* (Orthoptera), and *Locusta migratoria* (Orthoptera), but *A. domesticus* was over-represented in the metabarcoding datasets compared to the real abundance in terms of biomass. Fairly small insects like *Aphidius colemani*, *A. ervi*, *Encarsia formosa* (Hymenoptera), *Aphidoletes aphidimyza* (Diptera) and *Folsomia candida* (Collembola), were more abundant

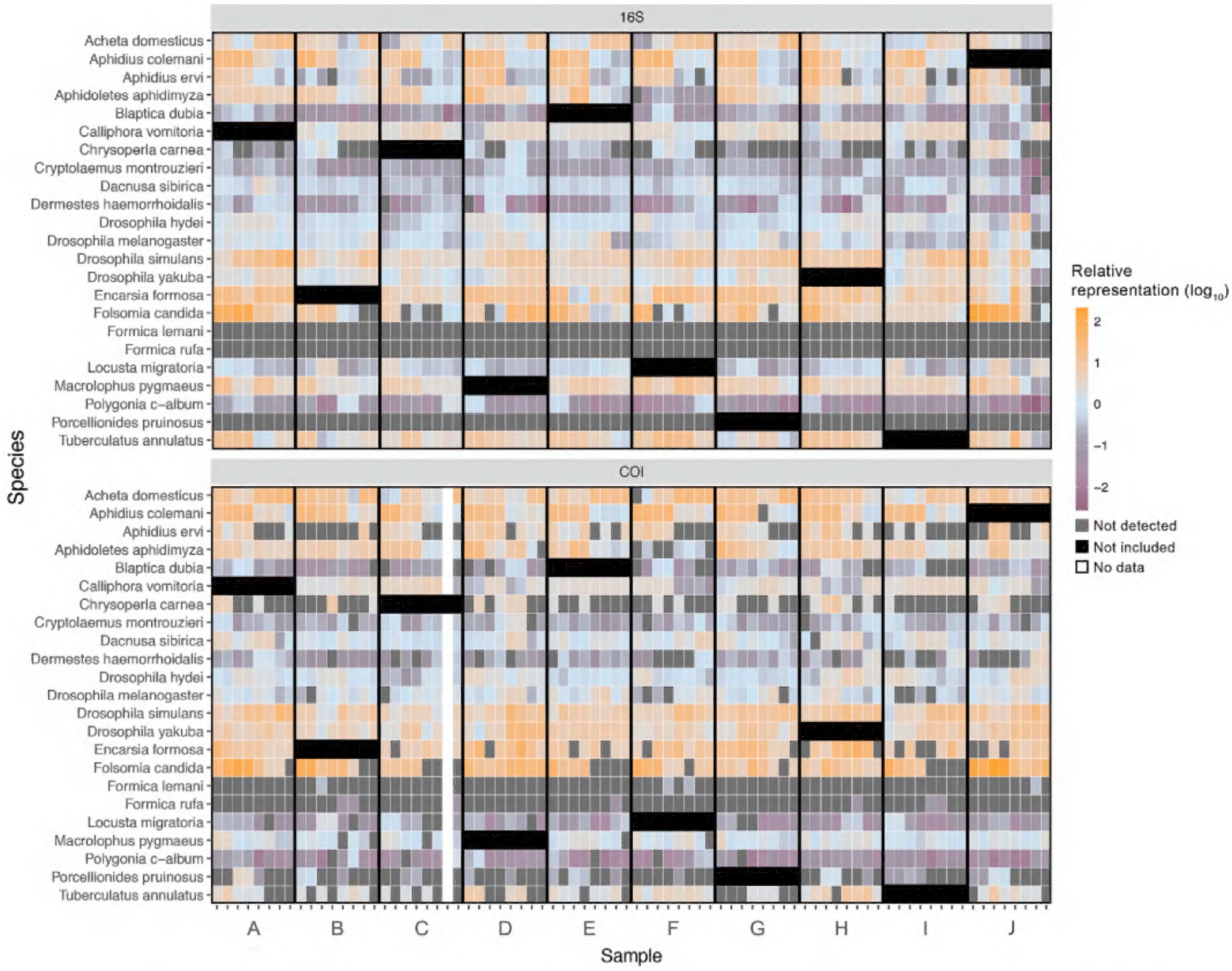

Figure 4. Representation of sequencing reads relative to biomass per sample. Relative representation is calculated as the log-ratio between relative read abundance and relative abundance in biomass of each species in each replicate. A higher log-ratio indicates that the species is over-represented in the metabarcoding dataset, while a lower value indicates that the species is under-represented in relation to its relative abundance in biomass in the mock community. Each community's replicates are indicated in the following order: B1.LT1.P1, B1.LT1.P2, B1.LT2.P1, B1.LT2.P2, B2.LT1.P1, B2.LT1.P2, B2.LT2.P1, and B2.LT2.P2. No reads were recovered for COI from sample C-B2.LT2.P1.

with respect to read numbers for both markers than in specimen number or biomass in the communities, as well as *Drosophila simulans* and *D. yakuba* (Diptera). On the other hand, larger species like *Blaptica dubia* (Blattodea), *Dermestes haemorrhoidalis* (Coleoptera), and *Polygonia c-album* (Lepidoptera), were also under-represented in the dataset. For other species, the bias depended on the marker. For instance, *Macrolophus pygmaeus* (Hemiptera) was over-represented in the 16S dataset, as well as some *Drosophila* species (Diptera) in the COI dataset, while *Cryptolaemus montrouzieri* (Coleoptera) was usually under-represented in the 16S dataset (Fig. 4, and Suppl. material 7: Figs S2, S3).

Regarding alpha diversity, replicates incubated in buffer B1 returned values of the Shannon Index (H') with a smaller decrease compared to those of the mock community based both on biomass and specimen numbers, irrespective of whether they had been incubated for LT1 or LT2 (Fig. 5, Suppl. material 7: Tables S7–S10). LT2 produced samples with a slightly lower value of H' in the 16S dataset, but the differences were not significant. In contrast, lysis in buffer B2 generated metabarcoding community estimates with a significantly lower diversity than the mock communities they originated from. This was true regardless of marker or lysis time.

The values of the Kullback-Leibler Divergences (i.e. the amount of information that is needed to transform the relative abundance distribution of species obtained with metabarcoding data into the original distribution of the mock communities) for the four treatments (two buffers, two incubation times) were quite similar regardless of whether community composition was based on biomass or specimen number (Fig. 6). For both 16S and COI, only Buffer had significant effects in the two cases (Suppl. material 7: Tables S11–S14). In the case of 16S, those replicates incubated in buffer B1 for LT1 had the lowest divergence values to the mock communities, although the interaction between Buffer and Lysis time was not significant, and a LT2 in B1 or incubation in buffer B2, increased the divergence. For COI, incubation time did not induce any difference in the value of the Kullback-Leibler Divergences, whereas the buffer in which the replicates were incubated did.

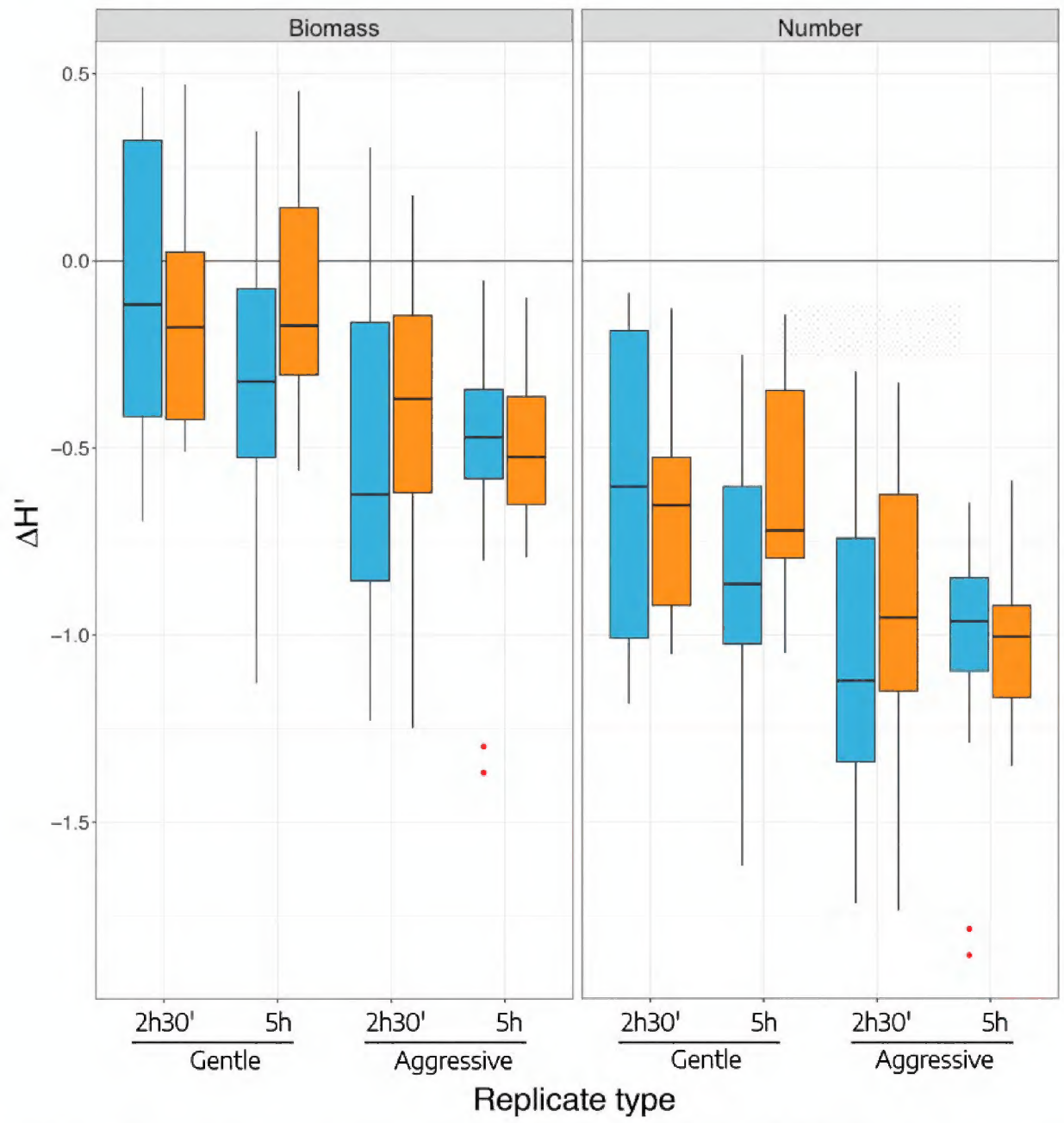

Figure 5. Estimated decrease in alpha diversity, measured as Shannon Index (H'), for different incubation treatments and markers. A short and gentle lysis (B1, LT1) recovers diversity values closer to the actual values of the original sample measured in biomass with the 16S marker (blue) and an increase in lysis time and chemical aggressiveness of the buffer returns more distant values. With COI (orange), this effect is dependent only on the lysis buffer. The decrease in alpha diversity compared to the mock samples based on number of individuals is greater than based on biomass, but they reproduce the same pattern. The black line indicates H'(mock community) = H'(metabarcoding sample).

Discussion

For single specimens, DNA extraction protocols that preserve the morphology of the insects have been used for more than a decade (Gilbert et al. 2007). Our results, however, add to only a handful of previous studies in demonstrating that mild digestion is good enough for extracting DNA from a bulk sample representing a mix of species. While many factors can still be manipulated to optimise these methods, our study reveals the explicit impact of three such factors – lysis buffer, digestion time and DNA purification method – on the recovery of known sample composition. In doing so, we carefully assess the quality of the insect material recovered after extraction.

Quality of morphological preservation

As far as we could judge, all lysis protocols applied here essentially left insect morphology intact. Much of this beneficial outcome may be due to the limited lysis time used, since the digestion step of each protocol here examined was less than 5 hours. The incubation times are, thus, much shorter than those used in other protocols for terrestrial insect samples, which range from around 14 hours of lysis to up to 72 hours (Vesterinen et al. 2016; Nielsen et al. 2019; Ji et al. 2020). The lysis times used here are more similar to those previously applied to aquatic invertebrates (Carew et al. 2018). In previous pilot tests (unpublished), we applied much longer digestion times of 48 – 72 hours, but then found them to be highly damaging for the insects, to an extent where identification even to order proved difficult and, in some cases, impossible.

The high level of morphological preservation here achieved is hope-inspiring. From a taxonomist perspective, it allows the later description of new species from the material treated. As an exciting scenario, we may then apply bulk metabarcoding to generate taxonomic lists of contents for large sets of bulk samples. Such lists may then be offered to expert taxonomists, allowing them to direct their input to those samples offering the highest reward in terms of new and interesting species to examine. This is a quantum leap from the tedious manual sorting of mass samples, where the main effort typically goes into dealing with the most abundant and typically less interesting taxa. Such tasks represent the poorest possible use of skilled taxonomists, whose availability tends to be in short supply.

DNA concentration and purity

All methods of DNA extraction tested here yielded DNA of sufficient concentration and quality for successful PCR and sequencing. DNA concentration was consistently higher for the replicates of which lysates had been purified with the salt saturation method, compared to those in which the extraction was done using silica-coated magnetic beads in a robot. This is not unexpected, as the salt saturation method started from a lysate volume of 7.5 mL, whereas the robot-based method started from only 225 µL, with both methods ending at a final elution volume of 150 µL. Thus, it is likely that the difference in initial quantity of DNA is reflected in the final concentration. In addition, it is important to note that the amount of beads in the reaction was the same for all four treatments, which might explain why the longer lysis times did not increase the DNA concentration more. In any case, the results are intuitive: a more chemically aggressive buffer, a longer digestion time and a larger input volume will all increase the concentration of DNA in the extract.

In terms of sample purity, the overall values achieved were high. Compared to the value of the A260/280 ratio considered ideal (1.8–2.0), we found highly adequate readings (1.5–1.9). In three of the four lysis treatments, the replicates purified with the salt saturation method produced a higher A260/A280 ratio, but those corresponding to buffer B2 and short incubation time showed the opposite relation. This could possibly be explained if one assumes that the higher concentration of proteinase K and the presence of DTT and Ca^{2+} in buffer B2 released more proteins to the lysate than buffer B1, but that the short incubation time was not enough to hydrolyse these proteins completely. However, in general, a longer incubation time produced DNA extracts with higher purity, same as the manual purification with the salt saturation method. Although significant, these differences had only a small effect on the species recovery. This differs from a previous study, in which the salt saturation method was shown to provide metabarcoding data with higher species richness than those provided by commercial kits(Kaunisto et al. 2017). However, the latter study was based on faecal material, and the results must, thus, be taken with caution. Since the contents of faeces depart from that of bulk insect sample, several other factors may bias the output (see, for example, Nielsen et al. 2018). Our results are in line with a more recent study (Nielsen et al. 2019) that used mock insect community samples and detected no differences in species recovery between the methods. As long as the lysate is well mixed, there is no indication that diversity recovery is affected by volume subsampling.

Accuracy in retrieving sample composition

In our study, all analyses were based on communities of known composition. In terms of species recovery, the COI marker was able to detect all 23 species we used in the communities, albeit not all species were detected in all the samples where they were present. For instance, small and delicate insects like *Tuberculatus annulatus* (Hemiptera) and *Chrysoperla carnea* larvae (Neuroptera)

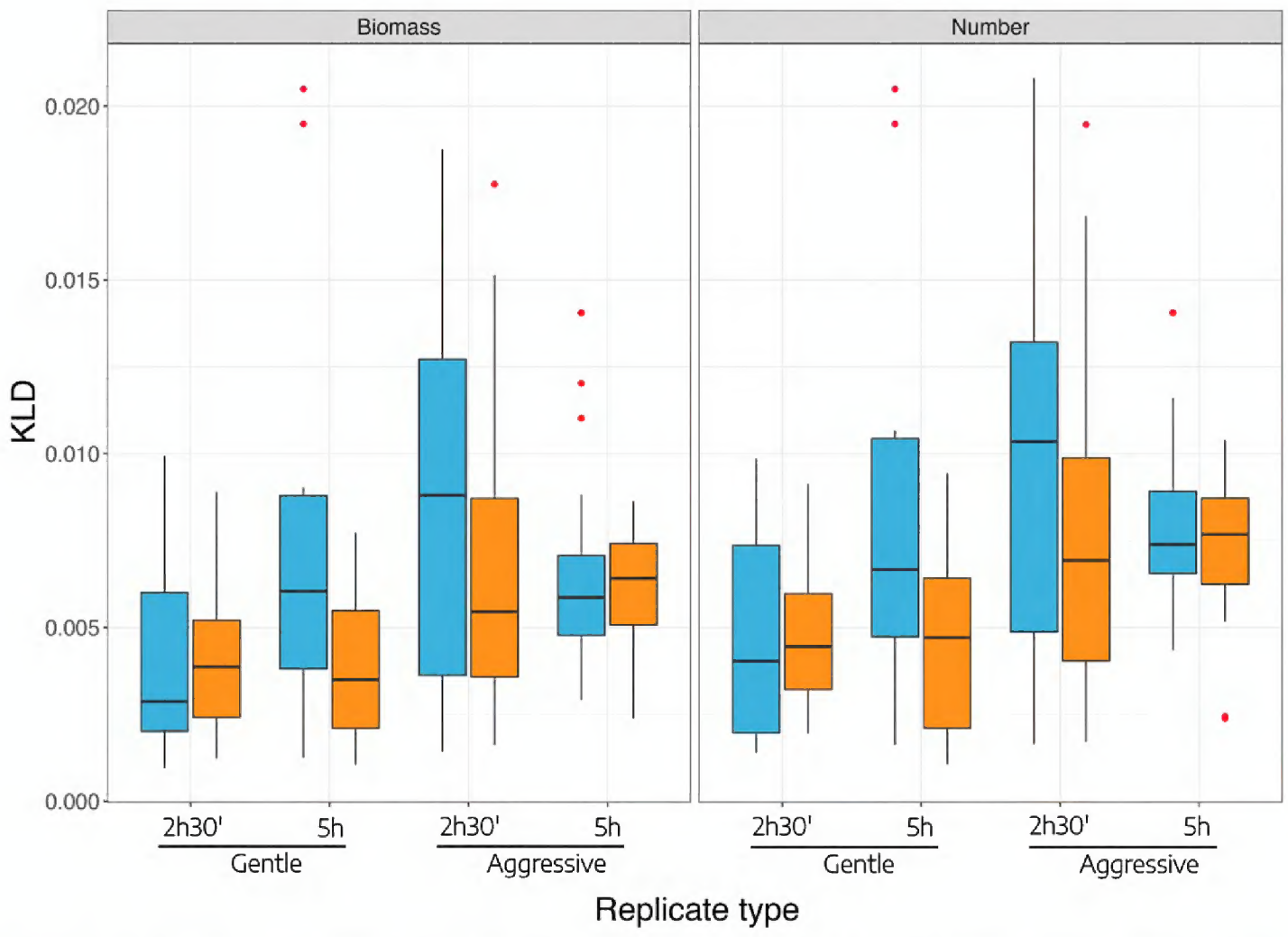

Figure 6. Kullback-Leibler Divergences between the true community composition and the metabarcoding estimates of it. Community composition is measured in terms of biomass (left) or the number of specimens (right). Data are shown both for the 16S marker (blue) and the COI marker (orange). For the 16S marker, the divergence between the metabarcoding and the original sample increases with buffer aggressiveness and incubation time, while for the COI marker, the divergence is only affected by an increase in buffer aggressiveness.

yielded low read abundances in most samples and appeared missing from many of those that were subjected to lysis with buffer B2. The 16S marker failed to detect the isopod *Porcellionides pruinosus* and the two species of *Formica* (Hymenoptera). The absence of *P. pruinosus* is not surprising, as the 16S primers used in this study had low degeneracy and were designed to target only insects (Marquina et al. 2019a). The case of *Formica* seems to be attributable to the resistance of the cuticle to mild lysis, as it is missing from the 16S dataset, but also very seldom recovered in the COI dataset, in combination with low binding affinity to both 16S and COI primers.

Importantly, the current study was explicitly aimed at evaluating the effect of the extraction protocol on the detectability of species against a community background of standardised complexity. Our communities were varied by excluding a single species amongst 23, whereas we did not vary the background complexity from highly species-poor to highly species-rich samples. As variation in the latter dimension provides an important aspect of natural communities, its effect should be the target of future studies. What we do see is that species detection rates, even against a standardised background, will never reach 100%. This pattern matches that reported by other studies. When using communities of known composition, both Carew et al. (2018) and Nielsen et al. (2019) reported significant variation in detection success, which they, too, attributed to a mix of effects in unknown proportions: species-specific characteristics, PCR biases and lack of homogeneity of the DNA extract. Thus, species-to-species variation in detectability remains an important challenge in the metabarcoding-based characterisation of insect communities, no matter which extraction method is used.

Future directions

Accurate abundance estimation is currently one of the main research fronts in metabarcoding. Early studies suggested that metabarcoding was unsuitable for quantification, and that accurate abundance estimates might only be achieved through shotgun sequencing using mitochondrial metagenomics (e.g. Crampton-Platt et al. 2016; Bista et al. 2018; reviewed in Lamb et al. 2019). Yet, recent developments using correction factors or spike-ins of known concentration (Krehenwinkel et al. 2017; Ji et al. 2020; Ershova et al. 2021) have yielded promising results. A complicating factor that will undoubtedly introduce additional noise when using mitochondrial markers to estimate abundance is that the mitogenome copy number may vary considerably between specimens due to size differences and to differences in physiological state (ratio of mitochondrion-rich to mitochondrion-poor tissue). However, this would affect both metagenomics and metabarcoding.

Importantly, the current study focuses on the impact of a single step in sample processing: that of the lysis phase of DNA extraction. What we find is some factors that clearly contribute to a poor overall relationship between species abundance and read abundance For instance, the failure of the metabarcoding data to estimate the specimen counts of the different species appears to be due to a large extent to the over-representation, at least in some treatments, of DNA from large species that were represented by only one or a few specimens (see, for example, *Calliphora*, *Acheta* and *Locusta* in Suppl. material 7: Figs S3, S4). It thus seems possible that the biases are consistent and that, given enough training data, a machine-learning model could compensate for the variance in read abundance for such taxa. If this proves the case, then it will be possible to improve the estimation of specimen counts considerably for other taxa.

In terms of other future improvements, it is quite plausible that mild lysis protocols can be further optimised to more accurately represent the contents of the sample processed. In our experiment, a moderately chemically aggressive lysis buffer and a short incubation time tended to reduce the difference in estimates of alpha diversity compared to the actual communities significantly more than a more destructive buffer or longer incubation times, showing that even when no precise estimates about species abundance distributions can be obtained, still some ecological insight can be drawn using this method. This likely illustrates simple considerations based on the relation between body volume and surface. During the early part of the lysis, both large and small insects presumably release DNA from their tissues in contact with the buffer, at a rate proportional to the exposed surface (roughly equivalent to the square of the body size). As the incubation time increases, the digestion will continue towards the internal tissues and, thus, the released DNA will be proportional to the volume of the individual (roughly equivalent to the cube of the body size) (Nielsen et al. 2019). Thus, it seems likely that larger insects will contribute proportionally more to the DNA pool the longer the lysis period or the more invasive the digestion buffer. These predicted patterns fit the observed Kullback-Leibler Divergences well, in that those replicates incubated in the gentler buffer and for a shorter period of time were more similar to the mock samples than those incubated in a more aggressive buffer or for a longer time. In this sense, longer incubation times would approximate the proportions of DNA from small versus large individuals obtained in homogenisation protocols, in which larger individuals contribute proportionally more to the DNA pool than smaller ones (Elbrecht et al. 2017b).

As a final caveat, we would like to re-emphasise that our results are based on a series of mock communities, the complexity of which is drastically lower than many real samples from Malaise traps or other efficient insect traps. In the future, experiments similar to the ones here conducted should thus be aimed at varying other aspects of community context, including significantly more diverse samples (Creedy et al. 2019), to verify that our results are consistent and scalable. Of course, the costs for the lysis will be higher for real Malaise trap samples (especially those from habitats with specimen-abundant insect faunas), but still affordable. We estimate that the current cost of lysis and DNA purification would be ~ 4.5 €/sample for a lysis volume

of 200 mL, and ~6 €/sample for a volume of 500 mL. The former volume will generally suffice for a typical Malaise trap sample, in our experience. For destructive sampling, extraction can be done for a smaller volume of biomass after homogenisation, thus lowering the reagent cost; on the other hand, the costs of equipment and tubes required for the homogenisation, or additional labour required, may not be trivial. However, we feel that any additional cost of mild lysis is more than outweighed by the benefits of being able to retain the specimens for later morphology-based work or single-specimen sequencing.

Conclusion

We have shown that non-destructive DNA extraction of mixed samples of terrestrial insects can provide DNA highly suitable for metabarcoding, while, at the same time, preserving the morphology of the individuals in good condition. Furthermore, our results indicate that a short and mild digestion followed by automated and commercially available DNA purification methods produces metabarcoding datasets that reliably retrieve most of the species in the original sample, while also providing closer approximations in measures linked to the relative abundance of the species. Metabarcoding can, thus, provide much help in the process of species discovery and description, and free up the expertise of taxonomists for the tasks where it matters the most.

Authors' Contributions

DM, TR, PL and FR conceived and designed the study; DM prepared the communities, conducted the experiment, analysed the data, prepared figures and tables and wrote the first draft of the manuscript. All authors contributed critically to subsequent versions of the manuscript and gave final approval for publication.

Data Availability

Raw sequencing reads assigned to samples can be accessed freely at https://zenodo.org/record/6559343#.YoYBR5NBw-R.

Acknowledgements

We are indebted to Brandon Cooper (University of Montana) and Laura Van Dijk (Stockholm University) for very generously providing us with numerous specimens of some of the species in our mock communities. We are also thankful to all members of the Ronquist lab and the team of the Insect Biome Atlas (https://www.insectbiomeatlas.com/) for the fruitful discussions during the design and analysis of the study. We thank Owen Wangensteen for his valuable comments on the manuscript. This project was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 642241 (BIG4 project, https://big4-project.eu) and by the Knut and Alice Wallenberg Foundation (KAW 2017.088). TR was further supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506; ERC-synergy project LIFEPLAN). Research reported in this publication was supported by the National Institute Of General Medical Sciences of the NIH of the US under award number R35GM124701 to Brandon S. Cooper. (insect cultures).

References

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Research 25(22): 4692–4693. https://doi.org/10.1093/nar/25.22.4692

Andrews S (2010) FastQC: A quality control tool for high-throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Batovska J, Piper AM, Valenzuela I, Cunningham JP, Blacket MJ (2021) Developing a non-destructive metabarcoding protocol for detection of pest insects in bulk trap catches. Scientific Reports 11(1): 1–14. https://doi.org/10.1038/s41598-021-85855-6

Beng KC, Tomlinson KW, Shen XH, Surget-Groba Y, Hughes AC, Corlett RT, Slik JF (2016) The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. Scientific Reports 6(1): e24965. https://doi.org/10.1038/srep24965

Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Molecular Ecology Resources 18(5): 1020–1034. https://doi.org/10.1111/1755-0998.12888

Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E (2016) obitools: A unix-inspired software package for DNA metabarcoding. Molecular Ecology Resources 16(1): 176–182. https://doi.org/10.1111/1755-0998.12428

Buchner D, Beermann AJ, Laini A, Rolauffs P, Vitecek S, Hering D, Leese F (2019) Analysis of 13,312 benthic invertebrate samples from German streams reveals minor deviations in ecological status class between abundance and presence/absence data. PLoS ONE 14(12): e0226547. https://doi.org/10.1371/journal.pone.0226547

Carew ME, Coleman RA, Hoffmann AA (2018) Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? PeerJ 6: e4980. https://doi.org/10.7717/peerj.4980

Cognato AI, Vogler AP (2001) Exploring data interaction and nucleotide alignment in a multiple gene analysis of Ips (Coleoptera: Scolytinae). Systematic Biology 50(6): 758–780. https://doi.org/10.1080/106351501753462803

Crampton-Platt A, Yu DW, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: Letting the genes out of the bottle. GigaScience 5(1): 15. https://doi.org/10.1186/s13742-016-0120-y

Creedy TJ, Ng WS, Vogler AP (2019) Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy. Ecology and Evolution 9(6): 3105–3116. https://doi.org/10.1002/ece3.4839

de Mendiburu F (2019) agricolae: statistical procedures for agricultural research. R package version 1.3-1. https://CRAN.R-project.org/package=agricolae

Elbrecht V, Leese F (2017a) Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. Frontiers of Environmental Science & Engineering in China 5: 314–311. https://doi.org/10.3389/fenvs.2017.00011

Elbrecht V, Peinert B, Leese F (2017b) Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. Ecology and Evolution 7(17): 6918–6926. https://doi.org/10.1002/ece3.3192

Erdozain M, Thompson DG, Porter TM, Kidd KA, Kreutzweiser DP, Sibley PK, Swystun T, Chartrand D, Hajibabaei M (2019) Metabarcoding of storage ethanol vs. conventional morphometric identification in relation to the use of stream macroinvertebrates as ecological indicators in forest management. Ecological Indicators 101: 173–184. https://doi.org/10.1016/j.ecolind.2019.01.014

Ershova EA, Wangensteen OS, Descoteaux R, Barth-Jensen C, Præbel K (2021) Metabarcoding as a quantitative tool for estimating biodiversity and relative biomass of marine zooplankton. ICES Journal of Marine Science 78(9): 3342–3355. https://doi.org/10.1093/icesjms/fsab171

Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8(1): 1–11. https://doi.org/10.1038/s41467-017-01312-x

Geller J, Meyer CP, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources 13(5): 851–861. https://doi.org/10.1111/1755-0998.12138

Gilbert MTP, Moore W, Melchior L, Worobey M (2007) DNA extraction from dry museum beetles without conferring external morphological damage. PLoS ONE 2(3): e272. https://doi.org/10.1371/journal.pone.0000272

Hajibabaei M, Spall JL, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC Ecology 12(1): 1–1. https://doi.org/10.1186/1472-6785-12-28

Hallmann CA, Sorg M, Jongejans E, Siepel H, Hofland N, Schwan H, Werner Stenmans W, Müller A, Sumser H, Hörren T, Goulson D, de Kroon H (2017) More than 75 percent decline over 27 years in total flying insect biomass in protected areas. PLoS ONE 12(10): e0185809. https://doi.org/10.1371/journal.pone.0185809

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America 101(41): 14812–14817. https://doi.org/10.1073/pnas.0406166101

Hebert PDN, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, Pedersen S, Jannetta P, DeWaard JR (2016) Counting animal species with DNA barcodes: Canadian insects. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 371(1702): e20150333. https://doi.org/10.1098/rstb.2015.0333

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang X, Levi T, Lott M, Emerson BC, Yu DW (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16(10): 1245–1257. https://doi.org/10.1111/ele.12162

Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW, Ovaskainen O (2020) SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Molecular Ecology Resources 20(1): 256–267. https://doi.org/10.1111/1755-0998.13057

Karlsson D, Hartop E, Forshage M, Jaschhof M, Ronquist F (2020) The Swedish Malaise Trap Project: A 15 year retrospective on a country-wide insect inventory. Biodiversity Data Journal 8: e47255. https://doi.org/10.3897/BDJ.8.e47255

Kaunisto KM, Roslin T, Sääksjärvi IE, Vesterinen EJ (2017) Pellets of proof: First glimpse of the dietary composition of adult odonates as revealed by metabarcoding of feces. Ecology and Evolution 7(20): 8588–8598. https://doi.org/10.1002/ece3.3404

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics (Oxford, England) 28(12): 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research 40(1): e3. https://doi.org/10.1093/nar/gkr771

Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports 7(1): 1–12. https://doi.org/10.1038/s41598-017-17333-x

Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI (2019) How quantitative is metabarcoding: A meta-analytical approach. Molecular Ecology 28(2): 420–430. https://doi.org/10.1111/mec.14920

Langor DW (2019) The diversity of terrestrial arthropods in Canada. ZooKeys 819: 9–40. https://doi.org/10.3897/zookeys.819.31947

Linard B, Arribas P, Andújar C, Crampton-Platt A, Vogler AP (2016) Lessons from genome skimming of arthropod-preserving ethanol. Molecular Ecology Resources 16(6): 1365–1377. https://doi.org/10.1111/1755-0998.12539

Liu M, Clarke LJ, Baker SC, Jordan GJ, Burridge CP (2019) A practical guide to DNA metabarcoding for entomological ecologists. Ecological entomology 6: e27295v2.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2015) Swarm v2: Highly-scalable and high-resolution amplicon clustering. PeerJ 3: e1420. https://doi.org/10.7717/peerj.1420

Marquina D, Andersson AF, Ronquist F (2019a) New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. Molecular Ecology Resources 19(1): 90–104. https://doi.org/10.1111/1755-0998.12942

Marquina D, Esparza-Salas R, Roslin T, Ronquist F (2019b) Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. Molecular Ecology Resources 19(6): 1516–1530. https://doi.org/10.1111/1755-0998.13071

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17(1): 10–12. https://doi.org/10.14806/ej.17.1.200

Martoni F, Nogarotto E, Piper AM, Mann R, Valenzuela I, Eow L, Rako L, Rodoni BC, Blacket MJ (2021) Propylene Glycol and Non-Destructive DNA Extractions Enable Preservation and Isolation of Insect and Hosted Bacterial DNA. Agriculture 11(1): 77. https://doi.org/10.3390/agriculture11010077

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How Many Species Are There on Earth and in the Ocean? PLoS Biology 9(8): e1001127. https://doi.org/10.1371/journal.pbio.1001127

Nielsen JM, Clare EL, Hayden B, Brett MT, Kratina P (2018) Diet tracing in ecology: Method comparison and selection. Methods in Ecology and Evolution 9(2): 278–291. https://doi.org/10.1111/2041-210X.12869

Nielsen M, Gilbert MTP, Pape T, Bohmann K (2019) A simplified DNA extraction protocol for unsorted bulk arthropod samples that maintains exoskeletal integrity. Environmental DNA 1(2): 144–145. https://doi.org/10.1002/edn3.16

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2013) Package Vegan: community ecology package, version 2.4-4. https://CRAN.R-project.org/package=vegan

Outhwaite CL, McCann P, Newbold T (2022) Agriculture and climate change are reshaping insect biodiversity worldwide. Nature 605(7908): 97–102. https://doi.org/10.1038/s41586-022-04644-x

Porter TM, Morris DM, Basiliko N, Hajibabaei M, Doucet D, Bowman S, Emilson EJS, Emilson CE, Chartrand D, Wainio-Keizer K, Guin ASX, Venier L (2019) Variations in terrestrial arthropod DNA metabarcoding methods recovers robust beta diversity but variable richness and site indicators. Scientific Reports 9(1): 1–11. https://doi.org/10.1038/s41598-019-54532-0

R Development Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: A versatile open source tool for metagenomics. PeerJ 4: e2584. https://doi.org/10.7717/peerj.2584

Ronquist F, Forshage M, Häggqvist S, Karlsson D, Hovmöller R, Bergsten J, Holston K, Britton T, Abenius J, Andersson B, Buhl PN, Coulianos C-C, Fjellberg A, Gertsson C-A, Hellqvist S, Jaschhof M, Kjærandsen J, Klopfstein S, Kobro S, Liston A, Meier R, Pollet M, Prous M, Riedel M, Roháček J, Schuppenhauer M, Stigenberg J, Struwe I, Taeger A, Ulefors S-O, Varga O, Withers P, Gärdenfors U (2019) Completing Linnaeus's inventory of the Swedish insect fauna: Only 5000 species left. PLoS ONE 15(3): e0228561. https://doi.org/10.1371/journal.pone.0228561

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Annals of the Entomological Society of America 87(6): 651–701. https://doi.org/10.1093/aesa/87.6.651

Srivathsan A, Hartop E, Puniamoorthy J, Lee WT, Kutty SN, Kurina O, Meier R (2019) Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. BMC Biology 17(1): 96. https://doi.org/10.1186/s12915-019-0706-9

Srivathsan A, Lee L, Katoh K, Hartop E, Kutty SN, Wong J, Yeo D, Meier R (2021) ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. BMC Biology 19(1): 1–21. https://doi.org/10.1186/s12915-021-01141-x

Statisticat LLC (2018) LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian-Inference.com. R package version 16.1.1. https://doi.org/10.1017/9781108646185.003

Taberlet P, Coissac E, Hajibabaei M, Riesenberg LH (2012) Environmental DNA. Molecular Ecology 21(8): 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Van Klink R, Bowler DE, Gongalsky KB, Swengel AB, Gentile A, Chase JM (2020) Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. Science 368(6489): 417–420. https://doi.org/10.1126/science.aax9931

Vesterinen EJ, Ruokolainen L, Wahlberg N, Peña C, Roslin T, Laine VN, Vasko V, Sääksjärvi IE, Norrdahl K, Lilley TM (2016) What you need is what you eat? Prey selection by the bat Myotis daubentonii. Molecular Ecology 25(7): 1581–1594. https://doi.org/10.1111/mec.13564

Zizka VMA, Leese F, Peinert B, Geiger MF (2019) DNA metabarcoding from sample fixative as a quick and voucher-preserving biodiversity assessment method. Genome 62(3): 122–136. https://doi.org/10.1139/gen-2018-0048

Supplementary material 1
Community types

Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist

Data type: excel file

Explanation note: Species composition of the ten different mock community types, with information on the abundance and biomass per species.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.78871.suppl1

Supplementary material 2
COI reference library

Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist

Data type: FASTA file

Explanation note: Reference fasta file for the COI barcodes of each species.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.78871.suppl2

Supplementary material 3
16S reference library
Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist
Data type: FASTA file
Explanation note: Reference fasta file for the 16S barcodes of each species.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.78871.suppl3

Supplementary material 4
Bioinformatic pipeline
Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist
Data type: pdf file
Explanation note: Detailed bioinformatic pipeline followed, specifying software used (with references), commands, and options for each step.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.78871.suppl4

Supplementary material 5
COI MOTU table
Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist
Data type: excel file
Explanation note: Metabarcoding dataset from the COI marker, with taxonomy of the species and reads/sample information.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.78871.suppl5

Supplementary material 6
16S MOTU table
Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist
Data type: excel file
Explanation note: Metabarcoding dataset from the 16S marker, with taxonomy of the species and reads/sample information.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.78871.suppl6

Supplementary material 7
Tables and figures
Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist
Data type: docx file
Explanation note: Supplementary tables (S1-S14) and figures (S1-S4).
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.78871.suppl7

Supplementary material 8
Lysis buffers and purification protocols
Author: Daniel Marquina, Tomas Roslin, Piotr Łukasik, Fredrik Ronquist
Data type: docx file
Explanation note: Detailed recipe for the two buffers used in the experiments and protocol for the manual purification method (P1).
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.78871.suppl8